Hindawi
Journal of Healthcare Engineering
Volume 2021, Article ID 6247652, 10 pages
https://doi.org/10.1155/2021/6247652

Research Article

Key Experimental Factors of Machine Learning-Based Identification of Surgery Cancellations

Fengyi Zhang,[1,2] **Xinyuan Cui**,[2,3] **Renrong Gong**,[4] **Chuan Zhang**,[5] **and Zhigao Liao**[1]

[1] *School of Management, Guangxi University of Science and Technology, Liuzhou, Guangxi Province 545006, China*
[2] *Business School, Sichuan University, Chengdu, Sichuan Province 610000, China*
[3] *School of Economics and Management, Harbin Institute of Technology, University Town, Nanshan, Shenzhen, China*
[4] *West China Hospital, Sichuan University, Chengdu, Sichuan Province 610000, China*
[5] *West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu, Sichuan Province 610000, China*

Correspondence should be addressed to Chuan Zhang; 331273145@qq.com

Received 6 October 2019; Revised 21 January 2020; Accepted 13 February 2021; Published 20 February 2021

Academic Editor: Kheng-Lim Goh

This study aimed to provide effective methods for the identification of surgeries with high cancellation risk based on machine learning models and analyze the key factors that affect the identification performance. The data covered the period from January 1, 2013, to December 31, 2014, at West China Hospital in China, which focus on elective urologic surgeries. All surgeries were scheduled one day in advance, and all cancellations were of institutional resource- and capacity-related types. Feature selection strategies, machine learning models, and sampling methods are the most discussed topic in general machine learning researches and have a direct impact on the performance of machine learning models. Hence, they were considered to systematically generate complete schemes in machine learning-based identification of surgery cancellations. The results proved the feasibility and robustness of identifying surgeries with high cancellation risk, with the considerable maximum of area under the curve (AUC) (0.7199) for random forest model with original sampling using backward selection strategy. In addition, one-side Delong test and sum of square error analysis were conducted to measure the effects of feature selection strategy, machine learning model, and sampling method on the identification of surgeries with high cancellation risk, and the selection of machine learning model was identified as the key factors that affect the identification of surgeries with high cancellation risk. This study offers methodology and insights for identifying the key experimental factors for identifying surgery cancellations, and it is helpful to further research on machine learning-based identification of surgeries with high cancellation risk.

1. Introduction

Surgery cancellation is a well-recognized health care quality problem that harms patients and wastes resources, leading to considerable losses for medical institutes and the entire health care system [1–3]. It forces scarce operative resources to remain idle and hinders patients' access to operative services [4]. Estimates of this revenue loss range between USD 1,430 and USD 1,700 for each cancelled case in US hospitals not on a fixed annual budget [5, 6]. In a review of surgery cancellations worldwide [4], the global cancellation rate (CR) generally ranges from 4.65% to 30.3%, which is a high proportion that urgently needs to be reduced. Identification of surgeries with high cancellation risk could provide information for health care service management and enable the adoption of preventive actions for achieving a lower CR [4]. Hence, it is of great value to identify surgeries with high cancellation risk.

For reflecting relevant information on patients and medical operation institutions, the hospital information system (HIS) plays an important role in health care service management, including surgery scheduling. Nowadays, health care service management-related studies on HISs have been conducted in many important fields [7–16], such as research on hospital admission rates [7, 8], clinical medication rules [9, 10], and referral management in hospitals [11, 12]. Particularly, the study of HIS in the field of identification of surgeries with high cancellation risk has proved that applying HIS data to identify surgeries with high and low risks of cancellation is feasible [4].

To date, most surgery cancellation studies have mainly focused on the predictors or risk factors of surgery cancellation. In the risk-factor-based category of cancellation, surgery cancellations generally include institutional resource-, capacity-, and patient-related cancellations. A significant percentage of surgery cancellations could be avoided, according to risk-factor-based analysis [17–19]. The risk factors of surgery cancellation include patient, admission, workup, and surgery schedule information, as well as administrative issues and surgery process records [4, 17, 20–26].

In recent years, various studies have focused on reducing surgery cancellation [21, 27–30]. For example, a team of researchers [27] used discrete event simulation modeling to represent perioperative processes and tested the effects of three scenarios on the number of surgical cancellations. Another team [28] simulated an anesthesiology preoperative assessment clinic to quantify the impact of patient information deficiency to mitigate the problem of surgery delay or cancellation. These studies used industrial engineering techniques to investigate means for reducing the number of surgical cancellations across the system but did not focus on identifying surgery with high cancellation risk. Among the works focusing on identifying surgeries with high cancellation risk, a retrospective cohort study [29] examined the association between patient, surgeon, and system factors and proved that several patient and system factors can be used to identify surgeries with a high likelihood of cancellation. The factors associated with surgery cancellation have been evaluated using chi-squared tests and multivariate logistic regression analyses [21]. Using multilevel logistic regression, an observational cohort study [30] identified patient- and hospital-level factors associated with cancellation owing to inadequate bed capacity. However, the accuracy of the above-mentioned studies only using traditional statistics models is low [29].

Machine learning is a powerful and effective tool for medical study. Machine learning has seen many applications in the fields of health care management [31–33], health care cost prediction [34, 35], and health care insurance [36–38]. Various machine learning models, which are of better performance compared with traditional statistics models, have been used in the field of identification of surgeries with high cancellation risk as well [4]. Particularly, Luo et al. [4] used sampling methods to handle the imbalance of the distribution of cancellation. Considering the performance of feature selection used in health care services [39, 40], Liu et al. [41] developed predictive models of last-minute surgery cancellation, in which forward selection was used as a feature selection method. However, there is no research that measured the effects of feature selection strategy, machine learning model, and sampling methods on the identification of surgery with high cancellation risk and identified the key factors of it.

This study aimed to provide effective methods for the identification of surgeries with high cancellation risk based on machine learning models and analyzed the key factors that affect the identification performance. One-side Delong test and sum of square error analysis were conducted to measure the effects of feature selection strategy, machine learning model, and sampling method on the identification of surgeries with high cancellation risk. This study offers methodology and insights for identifying the key experimental factors for identifying surgery cancellations, and it is useful to further research on machine learning-based identification of surgeries with high cancellation risk, in designing experimental process.

The rest of the paper is organized as follows. Section 2 provides detailed information about data collection and preparation, the methods used in this study, and model setup. The third part summarizes the experimental results. Section 4 further discusses the experiment and presents the findings. In Section 5, we draw conclusions from the findings and point out the direction of future work.

2. Data and Methods

2.1. Data. The data of this study were based on HIS sourced from West China Hospital, which is the largest hospital in southwest China. There are 70 operation rooms in West China Hospital nowadays, most of them usually open from 8 a.m. to 8 p.m., and daily average opening hours reach 12 hours. Overall, the data contained 5,125 cases from January 1, 2013, to December 31, 2014, of which 810 were cancelled (positive) and 4,315 were not, providing a CR of 15.80%. The hospital implemented the surgery day system. In this system, the surgeon has main surgery days; that is, the surgeon has the priority to use the operating room and its equipment on these days. All surgeries were scheduled one day in advance, and after confirmation, medical staff (surgeon, anesthesiologist, and nurse), the operation room, and the patient are bundled together. All cancellations were of institutional resource- and capacity-related types. Apart from the features originally obtained from HIS, we designed some features, according to the experience and knowledge of senior health care managers in West China Hospital, and they are whether the surgeon had surgery before (WSHSB), whether the surgery day was a legal holiday (WSDLH), whether it was the main surgery day (WMSD), whether surgeries have been cancelled (WC), and the number of days admitted (NDA). All the collected features of surgeries are listed in Table 1.

In the following experiment process, we considered not only the relationship between predictors and surgery cancellations but also the accessibility of predictors. As a result, 14 related predictors (features) were preliminarily refined, as described in Table 2, which covered all five information categories mentioned above. Among them, surgery schedule information contained five predictors: operating room (OR), surgeon, number of surgeries in the OR on a day (NSOD), the order number of surgery (ONS), and WSHSB. Then, patients' information and administrative issues contained three predictors each. For patients' information, they were age, sex, and anesthesia type (AT). For administrative issues, they were WSDLH, WMSD, and WC. Subsequently, workup and admission information contained two predictors and

Table 1: Predictors collected in this study.

Category	Number of predictors	Predictor(s)
Patients' information	4	Name, age, sex, and AT
Admission information	4	NDA, visit number, identification number of patient, and register number
Workup information	7	Drug allergy, names of drugs administered, blood type, WHSB, SN, and ST
Surgery schedule information	7	ONS, OR, surgery date, surgery time, surgeon, NSOD, WSHSB, and purpose of surgery
Administrative issues	10	Operation staff, department, ward, BD, last updated time, staff who last updated the information, WSDLH, WMSD, WC, and surgery expenditure
Surgery process records	23	Actual date/time when surgery began/ended, actual date/time when patient left OR, actual date when anesthesia was started, actual time when anesthesia was ended, actual date/time when predictive medicine was administered, body temperature, blood transfusion in surgery, autologous blood, allogeneic blood, plasma, thrombocyte, pathological examination, state of consciousness, general skin conditions, special skin conditions, drainage situation, surgery item delivery, anesthesia degree, and surgical incision category

AT: anesthesia type. BD: bed number. NDA: number of days admitted. NSOD: number of surgeries in the OR on the day. ONS: order number of surgery. OR: operating room. SN: surgery name. ST: surgery type. WC: whether surgery is cancelled. WHSB: whether there has been a surgery before. WMSD: whether it is the main surgery day. WSDLH: whether the surgery day is a legal holiday. WSHSB: whether the surgeon has surgery before.

Table 2: Predictors considered in this study.

Category	Number of predictors	Predictor(s)
Patients' information	3	Age, sex and AT
Admission information	1	NDA
Workup information	2	WHSB and ST
Surgery schedule information	5	OR, surgeon, NSOD, ONS, and WSHSB
Administrative issues	3	WSDLH, WMSD, and WC

AT: anesthesia type. NDA: number of days admitted. NSOD: number of surgeries in the OR on the day. ONS: order number of surgery. OR: operating room. ST: surgery type. WC: whether surgery is cancelled. WHSB: whether there has been a surgery before. WMSD: whether it is the main surgery day. WSDLH: whether the surgery day is a legal holiday. WSHSB: whether the surgeon has surgery before. Predictor(s): predictors mentioned above were preliminarily identified through expert interviews.

one predictor, respectively. Workup information contained WHSB and surgery type (ST), whereas admission information contained NDA.

2.2. Methods. The methods used in this study are introduced in the following three aspects: feature selection strategies, machine learning models, and sampling methods.

2.2.1. Machine Learning Models. Considering that the objective of this study is to identify surgery cancellation based on historical HIS data and different predictors, we modeled it as a supervised classification problem and utilized a representative set of machine learning models: random forest (RF), logistic regression (LR), extreme gradient boosting-tree (XGBoost-tree), support vector machine-linear (SVM-linear), and neural networks (NNET).

RF is a classifier composed of an ensemble of decision trees for training and predicting, which is widely used in medical management [42–44]. It is known for the high predictive performance and ability to find complex interactions among features [41, 45]. LR, a classification algorithm derived from linear regression, is a common approach employed in prediction and reasonable benchmark for evaluating other models. Compared with SVM and NNET, it has better interpretability that is important for model understanding and interpretation. XGBoost is an improved algorithm based on the gradient boosting decision tree with more detailed classification, XGB-Linear and XGB-tree. As for XGB-tree, it can construct boosted trees efficiently and when performing node splitting, the gain calculation of different predictors can be performed in parallel. SVM-linear is a kind of generalized linear classifier. Because of its advantages of solving high-dimensional pattern recognition problems and high accuracy, it is applied in this study. NNET is a model like the human brain's ability to predict and categorize, which learns the relationship between independent variables and dependent variables. Nowadays, it has been successfully applied in the classification and prediction of biological and medical data [46–48].

2.2.2. Feature Selection Strategies. This study takes four commonly used feature selection strategies into consideration, and they were forward selection strategy, backward selection strategy, LASSO-based strategy, and importance-based strategy. Forward selection starts with an empty set and iteratively adds the most important feature to the target feature subset from the candidate feature set, while backward selection iteratively removes the least important feature from the candidate feature set [49]. LASSO-based strategy refers to the cost function of the linear regression model added with the constraint of the L1 norm. It uses the control parameter for variable selection and complexity adjustment and is widely used in the medical field [50]. As for the importance-based strategy, RF, an excellent classifier model that has good applicability for feature selection [51, 52], was used to generate a dataset of the importance of each feature after training, and features with positive importance were considered as the most useful features to model training. Hence, feature selection using RF was conducted to select the useful predictors for the identification of the surgeries with high cancellation risk.

2.2.3. Sampling Methods. This study involves three sampling methods: oversampling, undersampling, and original sampling. Oversampling was intended to extract negative samples with replacement until the number of them was consistent with the number of positive samples. Conversely, undersampling extracted positive samples without replacement until their number was the same as the negative samples. Both methods change the class distribution of training data and have been used to address class imbalance [53]. These methods have performed well in several fields, such as in churning predictions related to bank credit cards [54] and classifying poor households [55]. Meanwhile, original sampling is the sampling of the original data set without making changes.

2.3. Experiment Setup. Figure 1 shows the entire experimental process mentioned below. For the preliminarily determined 14 related predictors, they may not certainly lead to the best performance of identification of surgery cancellation. Hence, the four feature selection strategies mentioned above (forward selection strategy, backward selection strategy, LASSO-based strategy, and importance-based strategy) were considered to achieve better performance. Subsequently, for a certain strategy, N predictors were obtained. To a large extent, the value of N is different for each strategy. Hence, predictors selected by original strategy (i.e., no feature selection is conducted) and four different feature selection strategies were applied to the following experiments, respectively.

All samples were divided into two sets, the train and test sets, at a ratio of 8 : 2. Based on this division, the train set was divided into the actual train and the validation sets, at a ratio of 6 : 2. Because of the imbalance in the positive-negative ratio (2 : 11) of the actual train set, we employed not only the original sampling but also over- and undersampling to achieve better performance. The train set and the actual train set were used to train the machine learning models; the validation set was used to determine the hyperparameters of models; and the test set was employed to validate the performance of the machine learning models. For each model, we performed a fivefold cross validation against the test set, nested within which was a fourfold cross validation against the validation set. Cross-validation methods are used to generate folds randomly, which refer to the combination of training and test data subset splits for training and validating machine learning models [56].

In this study, we designed 75 schemes (i.e., 5 × 5 × 3; five feature selection strategies, five machine learning models, and three sampling methods correspond to each other, as mentioned above), and each scheme was run to obtain the performance metrics of the test set. Fivefold cross validation was performed for each scheme to find and validate the optimal model with the best performance. Hence, 375 results of experiments were obtained eventually, from which the scheme with the best performance could be found and the influence of the factors (i.e., feature selection strategy, machine learning model, and sampling method) in surgery cancellation forecasting could be evaluated.

As the identification of surgeries with high cancellation risk in this study belongs to the binary classification problem, the performance of identification was measured according to seven metrics: accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), net reclassification index (NRI), and area under the curve (AUC) of the receiver operating characteristic (ROC). Sensitivity, specificity, PPV, and NPV are the metrics for model evaluation, which are used to reflect performance in a certain aspect. Sensitivity refers to the ratio of the correctly predicted positive sample number to the total number of true positive samples; conversely, specificity refers to the ratio of correctly predicted negative samples to the total number of true negative samples. Meanwhile, both PPV and NPV are metrics that focus on predictive samples. PPV refers to the ratio of the number of correctly predicted positive samples to the number of predicted positive samples, whereas NPV refers to the ratio of the number of correctly predicted negative samples to the number of predicted negative samples. Accuracy, NRI, and AUC are all used to reflect the overall performance of the model. Accuracy is the ratio of the correctly predicted sample number to the total predicted sample number. It does not distinguish the predicted sample as positive or negative. NRI is a measure of the change in risk prediction obtained when the risk marker under evaluation is added to an existing risk prediction model [57]. It was intended to serve as a summary measure to highlight the difference between two models [58]. AUC considers the imbalance of positive and negative samples and is often used with the ROC curve to illustrate performance assessments so that sensitivity and specificity can be considered in a comprehensive manner. In addition, feasibility is defined as the ability to make a considerable identification, and robustness refers to the measured performance close to the essential performance. In this study, AUC was considered to be the key metric; feasibility and robustness both focus on AUC.

The mentioned metrics were firstly analyzed for top schemes, and then we analyzed statistics (mean, maximum, minimum, etc.) of all schemes. Considering AUC being the key metric, we summarized the average AUCs grouped by different factors (feature selection strategy, machine learning model, and sampling method). We also conducted a Delong test [59] to evaluate the impact of different methods on predicting results. The variables involved in the study

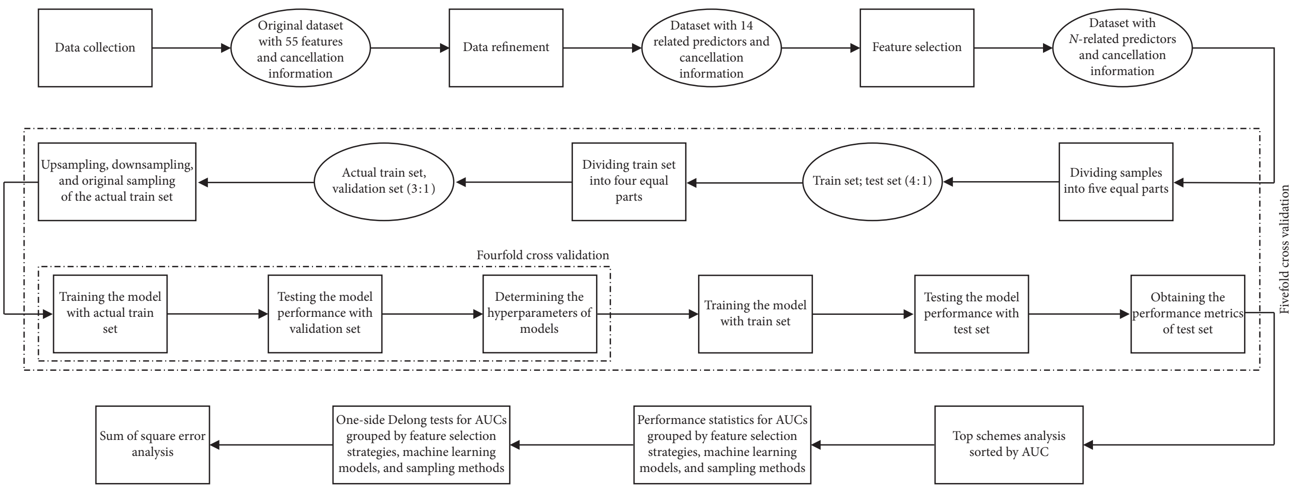
Data collection
Original dataset with 55 features and cancellation information
Data refinement
Dataset with 14 related predictors and cancellation information
Feature selection
Dataset with N-related predictors and cancellation information
Dividing samples into five equal parts
Train set; test set (4 : 1)
Dividing train set into four equal parts
Actual train set, validation set (3 : 1)
Upsampling, downsampling, and original sampling of the actual train set
Fourfold cross validation
Training the model with actual train set
Testing the model performance with validation set
Determining the hyperparameters of models
Training the model with train set
Testing the model performance with test set
Obtaining the performance metrics of test set
Fivefold cross validation
Top schemes analysis sorted by AUC
Performance statistics for AUCs grouped by feature selection strategies, machine learning models, and sampling methods
One-side Delong tests for AUCs grouped by feature selection strategies, machine learning models, and sampling methods
Sum of square error analysis

Figure 1: Experiment flowchart.

include feature selection methods, machine learning methods, sampling methods, and fivefold cross validation. Hence, the difference in AUC between specific methods can be evaluated by controlling other variables unchanged. $P < 0.05$ is considered statistically significant. In addition, the between-groups sum of squares (BGSS), within-groups sum of squares (WGSS), and total sum of squares (TSS) were used to measure deviation of the AUCs grouped by machine learning models, sampling methods, and feature selection strategies, and their definitions are given as follows:

$$\begin{aligned}\mathrm{BGSS} &= \sum_{i=1}^{r} n_i \left(\overline{x}_i - \overline{x}\right)^2,\\ \mathrm{WGSS} &= \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(\overline{x}_i - x_{ij}\right)^2,\\ \mathrm{TSS} &= \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(\overline{x} - x_{ij}\right)^2 = \mathrm{WGSS} + \mathrm{BGSS},\end{aligned} \tag{1}$$

where r refers to the number of groups, n_i refers to the number of examples belonging to group i, $\overline{x}$ refers to the mean of all samples, $\overline{x}_i$ refers to the mean of i^{th} group, and x_{ij} refers to the j^{th} sample of the i^{th} group. Given certain TSS, the larger the GBSS (the less the WGSS), the better the grouping. The experiments implemented were based on R software (version 3.61); the identification was conducted with the "caret" packages.

3. Results

We analyzed the experimental results in both scheme and factor level.

3.1. Analysis at Scheme Level. The schemes were mainly measured by averaging mentioned metrics of the fivefold cross validation, and the top 15 schemes in the test set are shown in Table 3, in descending order of average AUC. A model is considered as considerable predictive performance, if it is of a higher than 0.7 AUC [60, 61]. All NRIs were measured compared to the scheme of RF, backward selection strategy, and original sampling. Table 3 indicates the following: (1) The RF model with original sampling using backward selection strategy achieved the best performance according to accuracy (0.8578) and AUC (0.7199). (2) All top 9 schemes were of RF models, and RF model accounted for 11 of the top 15. Meanwhile, the schemes with a higher than 0.7 AUC were all RF models. (3) For the sampling method, original sampling and oversampling both accounted for 6 in the top 15 schemes, and the top 4 were original sampling. For the schemes with a higher than 0.7 AUC, both oversampling and original sampling accounted for half of all (4) For the NRI, there were 12 schemes with negative results, indicating that their performance was worse than the first one. Only two schemes achieved the positive NRI results, but differences were quite small (RF, original strategy, and original sampling: 0.0020; RF, importance-based strategy, and original sampling: 0.0057).

In addition, the statistics of each metric over the 75 schemes are shown in Table 4. According to Table 4, we can find the following: (1) for all the schemes, the specificity and NPV were quite high, with an average value of more than 0.75 (specificity: 0.8751, NPV: 0.7760), and the maximum value of NPV was 0.9988. In comparison, the sensitivity and PPV values were relatively small, whose mean values are less than 0.4 (sensitivity: 0.3215, PPV: 0.3938). (2) For each scheme, the standard deviations of AUC and specificity were small (AUC: 0.0430; specificity: 0.0208), which means that the values of them are stable.

3.2. Analysis at Factor Level. AUC is a useful metric that comprehensively reflects the performance of the model. The average AUCs grouped by each factor (feature selection strategy, machine learning model, and sampling method) are shown in Table 5. The average AUCs grouped by feature selection strategy indicate the following: (1) schemes using LASSO-based strategy had the largest AUC mean (mean: 0.6582), and these using forward selection strategy had the smallest (mean: 0.6426). However, the difference is quite

Table 3: Top 15 schemes in the test set.

	Model	Strategy	Sampling	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV	NRI
1	RF	Backward selection strategy	Original sampling	0.7199	0.8578	0.6712	0.8668	0.1963	0.9819	0.0000
2	RF	Original strategy	Original sampling	0.7136	0.8574	0.6740	0.8660	0.1901	0.9826	0.0020
3	RF	Forward selection strategy	Original sampling	0.7135	0.8560	0.6284	0.8692	0.2173	0.9759	−0.0404
4	RF	Importance-based strategy	Original sampling	0.7131	0.8574	0.6783	0.8654	0.1852	0.9835	0.0057
5	RF	Backward selection strategy	Oversampling	0.7055	0.8343	0.4579	0.8722	0.2654	0.9411	−0.2079
6	RF	Forward selection strategy	Oversampling	0.7034	0.8154	0.3904	0.8736	0.2963	0.9129	−0.2740
7	RF	Importance-based strategy	Oversampling	0.7029	0.8398	0.4869	0.8720	0.2580	0.9490	−0.1791
8	RF	Original strategy	Oversampling	0.7018	0.8394	0.4874	0.8726	0.2630	0.9476	−0.1780
9	RF	Forward selection strategy	Undersampling	0.6847	0.6176	0.2369	0.9003	0.6383	0.6137	−0.4008
10	NNET	LASSO-based strategy	Original sampling	0.6814	0.8398	0.4167	0.8424	0.0123	0.9934	−0.2789
11	RF	Backward selection strategy	Undersampling	0.6814	0.6328	0.2436	0.9009	0.6284	0.6336	−0.3935
12	RF	Original strategy	Undersampling	0.6808	0.6310	0.2384	0.8960	0.6074	0.6355	−0.4036
13	SVM-linear	LASSO-based strategy	Oversampling	0.6797	0.6597	0.2523	0.8970	0.5877	0.6732	−0.3887
14	LR	LASSO-based strategy	Oversampling	0.6794	0.6457	0.2493	0.9012	0.6185	0.6508	−0.3875
15	LR	LASSO-based strategy	Original sampling	0.6793	0.8429	0.5492	0.8459	0.0346	0.9947	−0.1429

LR: logistic regression model. NNET: neural networks. RF: random forest. SVM-linear: support vector machine-linear. XGBoost-tree: extreme gradient boosting-tree. Accuracy: ratio of the correctly predicted sample number to the total predicted sample number. Sensitivity: ratio of correctly predicted positive samples to the total number of true positive samples. Specificity: ratio of correctly predicted negative sample number to the total number of true negative samples. PPV: positive predictive value. NPV: negative predictive value. AUC: area under the receiver operating characteristic curve. NRI: net reclassification index; all NRIs were measured compared to the scheme of RF, backward selection strategy, and original sampling. The values in the table are averages of the fivefold cross validation.

Table 4: Statistics on the performance metrics.

	Max	Min	Mean	St. Dev.
AUC	0.7199	0.5310	0.6537	0.0430
Accuracy	0.8578	0.5967	0.7166	0.1016
Sensitivity	0.6783	0.1736	0.3215	0.1360
Specificity	0.9012	0.8419	0.8751	0.0208
PPV	0.6383	0.0031	0.3938	0.2362
NPV	0.9988	0.5914	0.7760	0.1634

Accuracy: ratio of the correctly predicted sample number to the total predicted sample number. Sensitivity: ratio of correctly predicted positive samples to the total number of true positive samples. Specificity: ratio of correctly predicted negative sample number to the total number of true negative samples. PPV: positive predictive value. NPV: negative predictive value. AUC: area under the receiver operating characteristic curve.

Table 5: Statistics on AUCs grouped by different factors.

Grouping factors	Groups	*N*	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max	Range
Feature selection strategy	Original strategy	75	0.6547	0.0428	0.5125	0.6372	0.6819	0.7292	0.2167
	LASSO-based strategy	75	0.6582	0.0441	0.4948	0.6449	0.6821	0.7239	0.2291
	Forward selection strategy	75	0.6426	0.0626	0.5085	0.6307	0.6846	0.7355	0.2270
	Backward selection strategy	75	0.6580	0.0427	0.5152	0.6435	0.6788	0.7331	0.2179
	Importance-based strategy	75	0.6550	0.0436	0.5125	0.6350	0.6823	0.7307	0.2182
Machine learning model	RF	75	0.6889	0.0314	0.6100	0.6713	0.7129	0.7355	0.1255
	LR	75	0.6666	0.0222	0.6297	0.6514	0.6792	0.7161	0.0864
	XGBoost-tree	75	0.6574	0.0256	0.6110	0.6353	0.6748	0.7171	0.1061
	NNET	75	0.6552	0.0259	0.6163	0.6362	0.6702	0.7239	0.1076
	SVM-linear	75	0.6005	0.0668	0.4948	0.5360	0.6558	0.7184	0.2236
Sampling method	Oversampling	125	0.6604	0.0403	0.5085	0.6385	0.6908	0.7355	0.2270
	Undersampling	125	0.6548	0.0354	0.5085	0.6398	0.6756	0.7239	0.2154
	Original sampling	125	0.6460	0.0626	0.4948	0.6365	0.6911	0.7340	0.2392

N: number of cases. Mean: mean value corresponding to AUC of each model. St. Dev.: standard deviation corresponding to AUC of each model. Pctl(25): AUC corresponds to the first quartile of the variance numerical distribution of each model. Pctl(75): AUC corresponds to the third quartile of the variance numerical distribution of each model. LR: logistic regression model. NNET: neural networks. RF: random forest. SVM-linear: support vector machine-linear. XGBoost-tree: extreme gradient boosting-tree.

Table 6: Superiority index of Delong test on AUCs grouped by feature selection strategies.

	Original strategy	LASSO-based strategy	Forward selection strategy	Backward selection strategy	Importance-based strategy
Original strategy	NA	23/75	8/75	8/75	24/75
LASSO-based strategy	13/75	NA	10/75	11/75	14/75
Forward selection strategy	11/75	21/75	NA	13/75	11/75
Backward selection strategy	0/75	18/75	5/75	NA	10/75
Importance-based strategy	11/75	21/75	2/75	10/75	NA

NA: not available.

small (0.0156). (2) In terms of standard deviations, except for the forward selection strategy (standard deviation: 0.0626), other strategies are of little difference between each other, and all of them were less than 0.0450. (3) The range (gap between maximum and minimum) of average AUC value of each strategy was around 0.2220.

The average AUCs grouped by machine learning model indicate the following: (1) RF model had the largest AUC mean (mean: 0.6889). (2) The ascending standard deviations were 0.0222 (LR), 0.0256 (XGBoost-tree), 0.0259 (NNET), 0.0314 (RF), and 0.0668 (SVM-linear models). (3) Compared with the other models, the average AUC of the RF model had the largest maximum (max: 0.7355).

The average AUCs grouped by sampling method indicate the following: (1) for oversampling and original sampling methods, AUC had the largest and the least mean values, respectively (oversampling: 0.6604; original sampling: 0.6460). (2) Among the three sampling methods, undersampling achieved the least standard deviation (standard deviation: 0.0354).

For comprehensively evaluating the performance of the factors mentioned above, one-sided Delong tests were conducted on the ROCs of each paired performance group; and the superiority index, which refers to the percentage of results of a certain group which precedes that of another group, is used to indicate the superiority between each factor. The superiority indexes are summarized in Tables 6–8 in the aspect of feature selection strategy, machine learning models, and sampling method, respectively, and only the factor-comparison groups with more than 1/2 superiority index were analyzed accordingly.

According to Tables 6 and 8, feature selection methods and sampling methods show no factor-comparison group with a more than 1/2 superiority index, which indicates that there is no factor (such as oversampling) superior to another factor (such as undersampling) with a threshold of 1/2. Table 7 presents the superiority index on machine learning models, and it indicates the following: (1) Compared with other machine learning methods, the superiority indexes of RF compared with other machine learning methods are all more than 1/2 (RF versus XGBoost-tree: 45/75, RF versus SVM-linear-tree: 56/75, and RF versus NNET: 49/75), except for the LR; however, the superiority index of RF compared to LR is 35/75, which is very close to 1/2 and much higher than that of LR compared to RF. The analysis above indicates the superiority of RF. (2) All superiority indexes of other machine learning methods compared to SVM-linear are more than 1/2, which indicates the inferiority of SVM-linear.

Table 7: Superiority index of Delong test on AUCs grouped by machine learning model.

	RF	LR	XGBoost-tree	SVM-linear	NNET
RF	NA	7/75	4/75	3/75	5/75
LR	35/75	NA	2/75	0/75	1/75
XGBoost-tree	45/75	7/75	NA	1/75	2/75
SVM-linear	56/75	48/75	46/75	NA	46/75
NNET	49/75	12/75	5/75	7/75	NA

LR: logistic regression model. NNET: neural networks. RF: random forest. SVM-linear: support vector machine-linear. XGBoost-tree: extreme gradient boosting-tree. NA: not available.

Table 8: Superiority index of Delong test on AUCs grouped by sampling method.

	Oversampling	Undersampling	Original sampling
Oversampling	NA	8/125	18/125
Undersampling	18/125	NA	25/125
Original sampling	20/125	18/125	NA

NA: not available.

Table 9 summarized the BGSS, WGSS, and TSS of AUCs grouped by three different factors mentioned above and indicated the following: (1) The maximum of BGSS (0.3191) and minimum of WGSS (0.5376) were obtained by machine learning models' grouping, which accounted for 37.25% and 62.75% of TSS, respectively. (2) The BGSS and WGSS grouped by sampling method were similar to the counterparts grouped by feature selection strategy (BGSS and WGSS grouped by sampling method: 0.0132 and 0.8435; BGSS and WGSS grouped by feature selection strategy: 0.0123 and 0.8443).

4. Discussion

The present results indicate the feasibility and robustness of identifying elective urologic surgeries with high cancellation risk. The average AUCs in the test set exceeded 0.65, with the maximum of AUC (0.7199, RF, original sampling, and backward selection strategy).

It is worth noting that RF model accounted for 11 of the top 15 among the 75 schemes, and the schemes with a higher than 0.7 AUC were all RF models. In addition, RF model was significantly different from the other models in the aspect of AUC. Finally, machine learning model factor achieved the maximum of BGSS and the minimum of WGSS. Hence, the

Table 9: Analysis of sum of square error.

	TSS	BGSS	BGSS/TSS (%)	WGSS	WGSS/TSS (%)
Grouped by feature selection strategy	0.8567	0.0123	1.44	0.8443	98.55
Grouped by ML model	0.8567	0.3191	37.25	0.5376	62.75
Grouped by sampling method	0.8567	0.0132	1.54	0.8435	98.46

TTS: total sum of square error. BGSS: between-groups sum of square error. WGSS: within-groups sum of square error.

selection of machine learning models is considered a key factor in identification of surgeries with high cancellation risk.

Among the top 8 schemes, the first four and the last four were original sampling and oversampling, respectively. However, sampling methods show no significant difference with each other according to the Delong test. In addition, there were large WGSS and small BGSS when grouped by sampling methods. In summary, sampling method is the factor that affects the identification of surgeries with high cancellation risk, but not as important as the machine learning methods, to some extent.

For different feature selection strategies, differences between their means were little and insignificant. In addition, there were large WGSS and small BGSS when grouped by sampling methods. Hence, feature selection strategy is not an important factor that affects the identification of surgeries with high cancellation risk.

This study has limitations. It only focused on the elective urologic surgeries in one hospital, which means that the data are single-centered. Although our predictors covered five information categories that previous studies have covered and on which good results have been obtained, there are also potentially useful predictors that have not been collected. Further research can consider conducting multicenter studies, including multiple hospitals and departments.

5. Conclusion

This study provided effective methods for the identification of surgeries with high cancellation risk based on machine learning models and analyzed the key factors that affect the identification of surgeries with high cancellation risk. It proved the feasibility and robustness of identifying surgeries with high cancellation risk, with the considerable maximum of AUC (0.7199) for RF with original sampling using a backward selection strategy. In addition, two-sided test and sum of square error analysis were conducted to measure the effects of feature selection strategy, machine learning model, and sampling method on the identification of surgeries with high cancellation risk; and the selection of machine learning model was identified as the key factors that affect the identification of surgeries with high cancellation risk. This study offers methodology and insights for identifying the key experimental factors for identifying surgery cancellations, and the insights of it are useful to further research on machine learning-based identification of surgeries with high cancellation risk, in designing experimental process.

Data Availability

The data supporting the study findings will not be shared since it is an organizational property. Data were anonymous, and study subjects could not be identified.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Fengyi Zhang conceived and designed the experiments; Fengyi Zhang performed the experiments; Renrong Gong contributed to reagents/materials/analysis tools; Xinyuan Cui, Chuan Zhang, and Zhigao Liao analyzed the data; Fengyi Zhang, Chuan Zhang, and Xinyuan Cui drafted or revised the manuscript; Fengyi Zhang and Xinyuan Cui approved the final version.

Acknowledgments

The work was supported in part by the National Natural Science Foundation of China (nos. 71532007, 71131006, and 71172197) and funded by the Sichuan University (2018skzx-pt173 and skzx2017-sb222) and the Innovation Project of Guangxi Graduate Education (YCSZ2014203).

References

[1] E. Hovlid, "A new pathway for elective surgery to reduce cancellation rates," *BMC Health Services Research*, vol. 12, 2012.

[2] V. Haana, K. Sethuraman, L. Stephens, H. Rosen, and J. G. Meara, "Case cancellations on the day of surgery: an investigation in an Australian paediatric hospital," *ANZ Journal of Surgery*, vol. 79, no. 9, pp. 636–640, 2009.

[3] E. Hovlid, "Patient experiences with interventions to reduce surgery cancellations: a qualitative study," *BMC Surgery*, vol. 13, 2013.

[4] L. Luo, F. Zhang, Y. Yao, R. Gong, M. Fu, and J. Xiao, "Machine learning for identification of surgeries with high risks of cancellation," *Health Informatics Journal*, vol. 26, no. 1, pp. 141–155, 2018.

[5] A. Macario, F. Dexter, and R. D. Traub, "Hospital profitability per hour of operating room time can vary among surgeons," *Anesthesia & Analgesia*, vol. 93, no. 3, pp. 669–675, 2001.

[6] F. Dexter, J. T. Blake, D. H. Penning, and D. A. Lubarsky, "Calculating a potential increase in hospital margin for elective surgery by changing operating room time allocations or increasing nursing staffing to permit completion of more

cases: a case study," *Anesthesia & Analgesia*, vol. 94, no. 1, pp. 138–142, 2002.

[7] A. M. Lavery, B. L. Banwell, G. Liu, and A. T. Waldman, "Hospital admission rates for pediatric multiple sclerosis in the United States using the pediatric health information system (PHIS)," *Multiple Sclerosis and Related Disorders*, vol. 9, pp. 5–10, 2016.

[8] D. Olson, J. K. Todd, M. Birkholz, E. J. Asturias, and J. T. Gaensbauer, "Analysis of the pediatric health information system database as a surveillance tool for travel-associated infectious diseases," *The American Journal of Tropical Medicine and Hygiene*, vol. 92, no. 5, pp. 1067–1069, 2015.

[9] C.-H. Jia, "Analysis of clinical medication rules in 48 398 patients with limb fractures based on hospital information system," *China Journal of Chinese Materia Medica*, vol. 41, no. 13, pp. 2555–2560, 2016.

[10] A. Atramont, D. Bonnet-Zamponi, I. Bourdel-Marchasson, I. Tangre, A. Fagot-Campagna, and P. Tuppin, "Health status and drug use 1 year before and 1 year after skilled nursing home admission during the first quarter of 2013 in France: a study based on the French national health insurance information system," *European Journal of Clinical Pharmacology*, vol. 74, no. 1, pp. 109–118, 2018.

[11] E. H. Kim, "Application and evaluation of personal health information management system," in *Proceedings of the 26th Annual International Conference of the Ieee Engineering in Medicine and Biology Society*, vol. 1–7, pp. 3159–3162, San Francisco, CA, USA, September 2004.

[12] M. Wang, C. Lau, F. A. Matsen, and Y. Kim, "Personal health information management system and its application in referral management," *IEEE Transactions on Information Technology in Biomedicine*, vol. 8, no. 3, pp. 287–297, 2004.

[13] S. A. Bittencourt, L. A. B. Camacho, and M. d. C. Leal, "O Sistema de Informação Hospitalar e sua aplicação na saúde coletiva," *Cadernos de Saúde Pública*, vol. 22, no. 1, pp. 19–30, 2006.

[14] H. Li, J. Wu, Y. Gao, and Y. Shi, "Examining individuals' adoption of healthcare wearable devices: an empirical study from privacy calculus perspective," *International Journal of Medical Informatics*, vol. 88, pp. 8–17, 2016.

[15] K. B. Eden, A. M. Totten, S. Z. Kassakian et al., "Barriers and facilitators to exchanging health information: a systematic review," *International Journal of Medical Informatics*, vol. 88, pp. 44–51, 2016.

[16] X. Wu, C. Wu, K. Zhang, and D. Wei, "Residents' numeric inputting error in computerized physician order entry prescription," *International Journal of Medical Informatics*, vol. 88, pp. 25–33, 2016.

[17] A. González-Arévalo, J. I. Gómez-Arnau, F. J. delaCruz et al., "Causes for cancellation of elective surgical procedures in a Spanish general hospital," *Anaesthesia*, vol. 64, no. 5, pp. 487–493, 2009.

[18] W. N. Schofield, G. L. Rubin, M. Piza et al., "Cancellation of operations on the day of intended surgery at a major Australian referral hospital," *Medical Journal of Australia*, vol. 182, no. 12, pp. 612–615, 2005.

[19] N. Al Talalwah and K. H. McIltrot, "Cancellation of surgeries: integrative review," *Journal of Perianesthesia Nursing*, vol. 34, no. 1, pp. 86–96, 2019.

[20] N. Sultan, A. Rashid, and S. M. Abbas, "Reasons for cancellation of elective cardiac surgery at prince sultan cardiac centre, Saudi Arabia," *Journal of the Saudi Heart Association*, vol. 24, no. 1, pp. 29–34, 2012.

[21] H. S. Cho, Y. S. Lee, S. G. Lee, J. M. Kim, and T. H. Kim, "Reasons for surgery cancellation in a general hospital: a 10-year study," *International Journal of Environmental Research and Public Health*, vol. 16, no. 1, 2019.

[22] K. Y. Yu, "Contributing factors of elective surgical case cancellation: a retrospective cross-sectional study at a single-site hospital," *BMC Surgery*, vol. 17, 2017.

[23] U. Yusuff and V. Handa, "Reasons for day of surgery cancellations: identifying trends to improve efficiency," *Anaesthesia*, vol. 72, p. 81, 2017.

[24] A. R. Seim, T. Fagerhaug, S. M. Ryen et al., "Causes of cancellations on the day of surgery at two major university hospitals," *Surgical Innovation*, vol. 16, no. 2, pp. 173–180, 2009.

[25] R. Kaddoum, "Causes of cancellations on the day of surgery at a tertiary teaching hospital," *BMC Health Services Research*, vol. 16, 2016.

[26] E. Turunen, M. Miettinen, L. Setälä, and K. Vehviläinen-Julkunen, "Elective surgery cancellations during the time between scheduling and operation," *Journal of Perianesthesia Nursing*, vol. 34, no. 1, pp. 97–107, 2019.

[27] S. Azari-Rad, A. Yontef, D. Aleman, and D. Urbach, "Reducing elective general surgery cancellations at a Canadian hospital," *Canadian Journal of Surgery*, vol. 56, no. 2, pp. 113–118, 2013.

[28] D. Morrice, "A simulation analysis of a patient-centered surgical home to improve outpatient surgical processes of care and outcomes," in *Proceedings of the 2013 Winter Simulation Conference*, pp. 2274–2286, Washington, DC, USA, December 2013.

[29] A. L. Tan, C. J. Chiew, S. Wang et al., "Risk factors and reasons for cancellation within 24 h of scheduled elective surgery in an academic medical centre: a cohort study," *International Journal of Surgery*, vol. 66, pp. 72–78, 2019.

[30] D. J. N. Wong, S. K. Harris, and S. R. Moonesinghe, "Cancelled operations: a 7-day cohort study of planned adult inpatient surgery in 245 UK national health service hospitals," *British Journal of Anaesthesia*, vol. 121, no. 4, pp. 730–738, 2018.

[31] T. A. Daghistani, R. Elshawi, S. Sakr, A. M. Ahmed, A. Al-Thwayee, and M. H. Al-Mallah, "Predictors of in-hospital length of stay among cardiac patients: a machine learning approach," *International Journal of Cardiology*, vol. 288, pp. 140–147, 2019.

[32] B. V. Nithya, "Predictive analytics in health care using machine learning tools and techniques," in *Proceedings of the 2017 International Conference on Intelligent Computing and Control Systems*, pp. 492–499, Madurai, India, June 2017.

[33] O. Kocsis, "Assessing machine learning algorithms for self-management of asthma," in *Proceedings of the 2017 IEEE International Conference on E-Health and Bioengineering Conference*, pp. 571–574, Sinaia, Romania, June 2017.

[34] D. Bertsimas, M. V. Bjarnadóttir, M. A. Kane et al., "Algorithmic prediction of health-care costs," *Operations Research*, vol. 56, no. 6, pp. 1382–1392, 2008.

[35] J. M. Karnuta, S. M. Navarro, H. S. Haeberle et al., "Predicting inpatient payments prior to lower extremity arthroplasty using deep learning: which model architecture is best?" *The Journal of Arthroplasty*, vol. 34, no. 10, pp. 2235–2241, 2019.

[36] I. Kose, M. Gokturk, and K. Kilic, "An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance," *Applied Soft Computing*, vol. 36, pp. 283–299, 2015.

[37] J. O. Seo, "Identifying frauds and anomalies in medicare-B dataset," in *Proceedings of the 2017 39th Annual International Conference of the Ieee Engineering in Medicine and Biology Society*, pp. 3664–3667, Jeju Island, Republic of Korea, July 2017.

[38] A. Chimmad, P. Saripalli, and V. Tirumala, "Assessment of healthcare claims rejection risk using machine learning," in *Proceedings of the 2017 IEEE 19th International Conference on E-Health Networking, Applications and Services (Healthcom)*, pp. 1–6, Dalian, China, October 2017.

[39] M. Swangnetr and D. B. Kaber, "Emotional state classification in patient-robot interaction using wavelet analysis and statistics-based feature selection," *Ieee Transactions on Human-Machine Systems*, vol. 43, no. 1, pp. 63–75, 2013.

[40] K. H. Kim, B. Ku, N. Kang, Y. S. Kim, J. S. Jang, and J. Y. Kim, "Study of a vocal feature selection method and vocal properties for discriminating four constitution types," *Evidence-Based Complementary and Alternative Medicine*, vol. 2012, Article ID 831543, 10 pages, 2012.

[41] L. Liu, Y. Ni, N. Zhang, and J. Pratap, "Mining patient-specific and contextual data with machine learning technologies to predict cancellation of children's surgery," *International Journal of Medical Informatics*, vol. 129, pp. 234–241, 2019.

[42] A. S. Fathima and D. Manimeglai, "Analysis of significant factors for dengue infection prognosis using the random forest classifier," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 2, pp. 240–245, 2015.

[43] H. Byeon, "Developing a random forest classifier for predicting the depression and managing the health of caregivers supporting patients with Alzheimer's Disease," *Technology and Health Care*, vol. 27, no. 5, pp. 531–544, 2019.

[44] V. Bellini, "Artificial intelligence: a new tool in operating room management. Role of machine learning models in operating room optimization," *Journal of Medical Systems*, vol. 44, no. 1, 2019.

[45] Z. ShahabiKargar, "Predicting procedure duration to improve scheduling of elective surgery," in *Pricai 2014: Trends in Artificial Intelligence*, D. N. Pham and S. B. Park, Eds., pp. 998–1009, Springer, Berlin, Germany, 2014.

[46] F. Jaimes, J. Farbiarz, D. Alvarez, and C. Martínez, "Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room," *Critical Care*, vol. 9, no. 2, pp. R150–R156, 2005.

[47] G. Clermont, D. C. Angus, S. M. DiRusso, M. Griffin, and W. T. Linde-Zwirble, "Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models," *Critical Care Medicine*, vol. 29, no. 2, pp. 291–296, 2001.

[48] G. Meyfroidt, F. Güiza, J. Ramon, and M. Bruynooghe, "Machine learning techniques to examine large patient databases," *Best Practice & Research Clinical Anaesthesiology*, vol. 23, no. 1, pp. 127–143, 2009.

[49] Z. M. He, "Quantum-enhanced feature selection with forward selection and backward elimination," *Quantum Information Processing*, vol. 17, no. 7, 2018.

[50] A. Marelli, "Machine learning developed algorithms to identify patients with congenital heart disease from large claims databases," *Circulation*, vol. 140, no. 1, 2019.

[51] N. T. Huang, "Feature selection of power quality disturbance signals with an entropy-importance-based random forest," *Entropy*, vol. 18, no. 2, 2016.

[52] H. Wang, "Using random forests to select optimal input variables for short-term wind speed forecasting models," *Energies*, vol. 10, no. 10, 2017.

[53] D. H. Shi, J. Guan, and J. Zurada, "Cost-sensitive learning for imbalanced bad debt datasets in healthcare industry," in *Proceedings of the 2015 Asia-Pacific Conference on Computer-Aided System Engineering—Apcase 2015*, C. Monsalve and A. Sanchez, Eds., pp. 30–35, Quito, Ecuador, July 2015.

[54] N. Naveen, V. Ravi, and D. A. Kumar, "Application of fuzzyARTMAP for churn prediction in bank credit cards," *International Journal of Information and Decision Sciences*, vol. 1, no. 4, pp. 428–444, 2009.

[55] B. Santoso, "A comparative study of synthetic over-sampling method to improve the classification of poor households in yogyakarta province," *IOP Conference Series: Earth and Environmental Science*, vol. 187, Article ID 012048, 2018.

[56] L. Feng, "Analysis of the influencing factors of elective surgery cancellation," in *Proceedings of the 2015 7th International Conference on Information Technology in Medicine and Education*, S. Li, Y. Dai, and Y. Cheng, Eds., pp. 41–45, Huangshan, China, November 2015.

[57] S. K. Zalawadiya, P. C. Gunasekaran, C. P. Bavishi, V. Veeranna, S. Panaich, and L. Afonso, "Left ventricular hypertrophy and risk reclassification for coronary events in multi-ethnic adults," *European Journal of Preventive Cardiology*, vol. 22, no. 5, pp. 673–679, 2015.

[58] L. E. Thomas, E. C. O'Brien, J. P. Piccini, R. B. D'Agostino, and M. J. Pencina, "Application of net reclassification index to non-nested and point-based risk prediction models: a review," *European Heart Journal*, vol. 40, no. 23, pp. 1880–1887, 2019.

[59] D. N. Jayasekara and M. R. Sooriyarachchi, "A simulation based study for comparing tests associated with receiver operating characteristic (ROC) curves," *Communications in Statistics—Simulation and Computation*, vol. 43, no. 10, pp. 2444–2467, 2014.

[60] X. Chen, "Prediction of optimal gene functions for osteosarcoma using gene ontology and microarray profiles," *Journal of Bone Oncology*, vol. 7, pp. 18–22, 2017.

[61] J. Yang, "Differential diagnosis of pancreatic serous cystadenoma and mucinous cystadenoma: utility of textural features in combination with morphological characteristics," *BMC Cancer*, vol. 19, no. 1, p. 1223, 2019.